International Journal of Mathematics and
Computer Applications Research (IJMCAR)
ISSN (P): 2249-6955; ISSN (E): 2249-8060
Vol. 9, Issue 2, Dec 2019, 29-38

TRANS STELLAR
•Journal Publications •Research Consultancy

LIVER FUNCTIONING: COMPARISON BETWEEN DIFFERENT CLASSIFICATION METHODS

VARDHANA NAVELKER, GAURI DANGUI, SHUBHECHHA BHARNE, VALERIE MENEZES & ASHWINA TARI

Department of Computer Engineering, Agnel Institute of Technology and Design, Assagao-Goa, India

ABSTRACT

The liver is one of the most crucial organs in the human body. In our project, we will be working to predict, based on various factors, whether the liver is functioning properly or not. This determination is conventionally done using a blood test, for verification by a physician against certain standard values of factors in the blood. There are 8 attributes (or factors) that the blood is tested for. These attributes include the following: Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, and Albumin to Globulin Ratio. There may be many more attributes, but they are not as crucial as the above-mentioned eight. A physician then determines the healthiness of the liver with the result of the blood test, by checking them against the standard normal range of the blood factors.

The aim of this project is to correctly determine whether the liver is functioning correctly or not i. e. if the liver is healthy or not. Hence, in this project, various classification methods will be used to determine the above. The dataset obtained has eight factors, along with the class label. The class label for the dataset is given by two discrete values: 1 & 2. To improve the accuracy of the classifier methods, we use the wrapper approach for choosing the best attribute subset, i. e., selecting a set of attributes that will give the most accurate result from amongst all attributes.

KEYWORDS: DGW, Wrapper Approach, Classifiers & Attribute Evaluator Rankers

Received: Jun 02, 2019; **Accepted:** Jun 22, 2019; **Published:** Jul 09, 2019; **Paper Id.:** IJMCARDEC20193

Original Article

1. INTRODUCTION

Liver (liver cells) perform/(s) over 400 different functions, along with other organs and systems. Liver function tests are blood tests to check how well the liver is working. Usually this determination of whether the liver is functioning well is done using a blood test, which is then verified by the physician against standard values or range of values of certain attributes in the blood tests. Our project's aim is to correctly determine whether or not the liver is working correctly, that is whether it is carrying out all the processes it is supposed to and, in the manner it is supposed to.

Considering the research done by us, on the functioning of the liver, our proposed system will determine whether the liver is functioning properly or not using the least number of attributes needed from the blood test results and improve the accuracy of the output that is if the liver is functioning properly or not.

2. PROPOSED SYSTEM

2.1 Structure of Proposed System

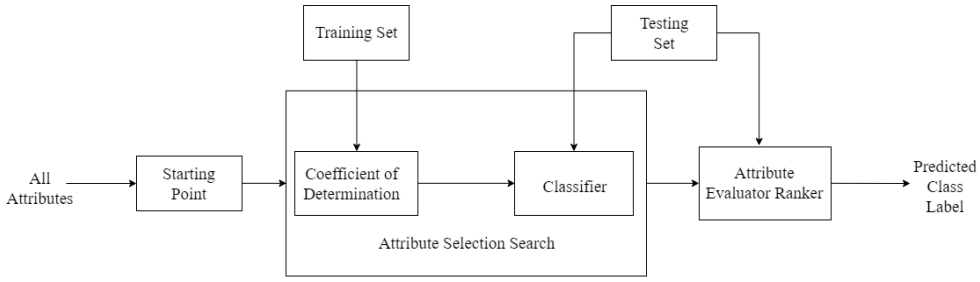

Figure 1: Flow of the Solution

2.1.1 Starting Point

The DGW Wrapper starting point contains all the original attributes. For this use case, we have all the eight factors in the set A, the set of all the original attributes. The attributes present in the ILPD include Age, Gender, TB, DB, ALP, ALT, AST, TP, ALB, A/G Ratio and selector. The class label would be crucial to tell liver patients and non-liver patients apart. The ILPD contains 416 liver patient records and 167 non-liver patient records. The liver patients are indicated using class label 1 and non-liver patients by class label 2.

2.1.2 Classifier

The classifier methods to be applied are chosen at this point in the DGW. In our proposed work, two classification algorithms are implemented: Iterative Dichotomizer 3 and Naïve Bayes classifier.

2.1.3 Attribute Selection Search

In this step, a two-stage search is performed:

- The first stage is Dependency Based Wrapping. To avoid all dependencies present on the original attributes, the Coefficient of Determination (r^2) is calculated for all attribute pairs. In the case of the ILPD, the COD for all possible combinations of the factors in blood are calculated and arranged in decrementing order, according to their values. There were 28 possible combinations of attributes
- In the next stage, the one pair of attributes is removed from the dataset. Then the classifier is applied. Evaluation measures like accuracy, error rate and precision are determined. The same procedure is repeated for all other attribute pairs.

2.1.4 Attribute Evaluator Ranker

Now for each attribute that remains in the best subset, an attribute evaluator is applied. For this particular use case we consider five different rankers:

- Gini Index
- Information Gain
- Gain Ratio

- One R

3. PREPROCESSING AND CLASSIFICATION

3.1. Preprocessing: Discretization

Since the dataset consisted of continuous attribute values, in order to apply the classifiers, it had to be converted to ordinal values. This was done using four methods:

- Brute Force Discretization
- Normal Range Discretization
- Adjacent Value Discretization
- Mean Value Discretization

On comparing the above four methods, using information gain, on a sample dataset, we derived the results as in figure 2. The best method was brute force discretization, hence it was used to discretise the full dataset for the system and the classifiers were applied on the resultant output.

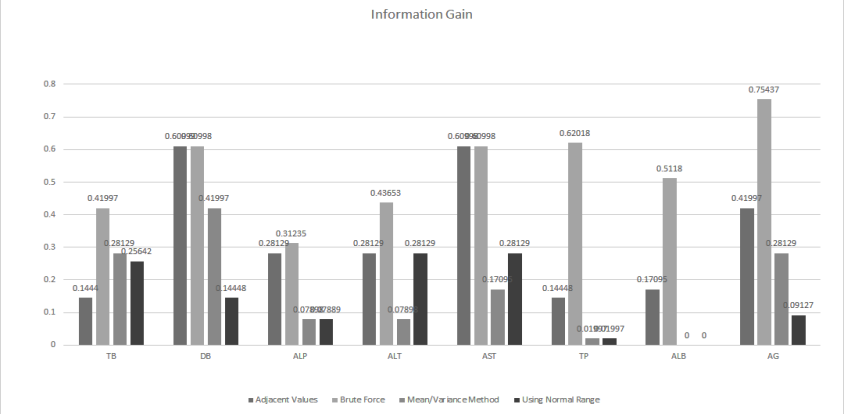

Figure 2: Comparison of IG using four Methods Mentioned Above

3.2. Classification

3.2.1 Iterative Dichotomiser 3

Iterative Dichotomiser 3 is a type of decision tree, with elements the same as in the decision tree, which are the internal nodes, branch and the leaf nodes.

Procedure:

- Calculate the entropies of all the attributes.
- Choose the attribute that has the lowest entropy or with the highest gain.
- Create a root node for the tree with the attribute selected in step 2. The child nodes will be the attribute split on its ordinal values.
- Repeat step 1 & 2 until the leaf nodes are pure, i. e. the class label for all entries is same.

Entropy = $\sum_{i=1}^{i} -p_i * \log_2 p_i$

Information Gain

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

A major problem in this classifier is that of further classifying the leaf node. We can encounter that the leaf nodes can no further be split. Then the count for each class label in the node is done, and the class label with the highest count is assigned as the label for the node. This procedure is majority voting.

3.2.2 Naïve Bayes Classifier

Naïve Bayesian is a statistical classifier. It predicts class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classifier is based on Bayes Theorem.

Procedure:

- Select an attribute from the list of attributes. Find the probability for each ordinal value of the attribute for each class label by using Bayes Theorem.
- Repeat step 1 for all attributes.
- Assume a class label. Take the test case and choose probabilities according to the ordinal values given the assumed level. Multiply the probability to find the P(1|Test Values).
- Repeat step 3 with the assumption of the other class label i. e. P(2|Test Values).
- The predicted label will be the higher value amongst the one calculated for step 3 & 4.

Bayes Theorem

$$P(X|Y) = \frac{P(Y|X) * P(X)}{P(Y)}$$

The major issue faced when applying this classifier is when all class labels have the same probability for a given test record. This issue can be solved by using Laplacian correction wherein all the class label counts are increased by one and the total count is increased by the number of classes.

4. RESULTS

4.1. Preprocessing

Since according to the results of the mathematical model, the best method for discretization was Brute Force Method. Hence, Brute Force Discretization was applied on the training set to get the split values. Using the split values, the test set was discretized at well. The split values along with their respective information gain values are as listed below.

Table 1: Attribute, Split Values, Information Gain

Attributes	Split Value	Information Gain
TB	1.6	0.08881
DB	1.2	0.08982
ALP	211	0.05764
ALT	61	0.05778
AST	47	0.06617
TP	3.6	0.00538

Table 1: Contd.,		
ALB	3.4	0.02339
AG	0.89	0.02528

4.2. Attribute Selection Search

4.2.1 Coefficient of Determination

The Coefficient of determination was calculated as proposed in the mathematical model. The values were the highest value of COD was found to be for TB-DB, followed by TB-ALB, ALT-AST and ALB-AG. The least value was for DB-TP, followed by TB-TP, ALT-AG and ALT-TP.

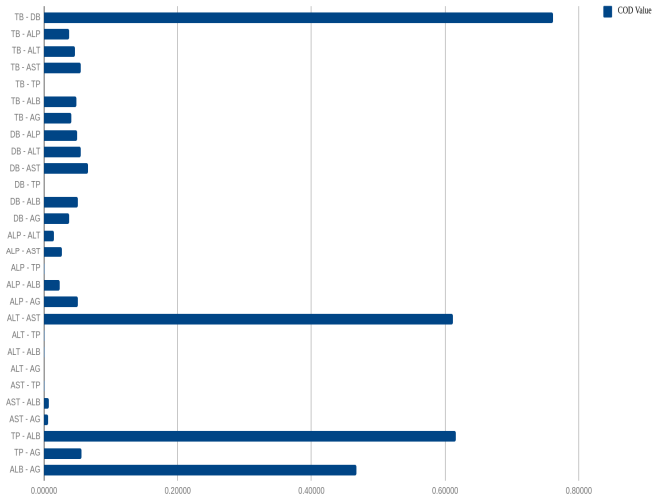

Figure 3: COD for all Attribute Pairs

4.2.2 Classifier: Iterative Dichotomizer 3

The classifier was applied when all attribute pairs were present to get the following results:

Table 2: Confusion Matrix and Evaluation Measures for Iterative Dichotomiser 3

Positive		Negative		Error Rate	Accuracy	Sensitivity	Specificity	Precision
True	False	True	False					
22	10	3	10	0.4444	0.5556	0.6875	0.2308	0.6875

4.2.3 Classifier: Naïve Bayes Classifier

The classifier was applied when all attribute pairs were present to get the following results:

Table 3: Confusion Matrix and Evaluation Measures for Naïve Bayes Classifier

Positive		Negative		Error Rate	Accuracy	Sensitivity	Specificity	Precision
True	False	True	False					
9	4	9	23	0.6	0.4	0.6923	0.2813	0.6923

The comparison of Coefficient of Determination (COD) values and the accuracies when attribute pairs are removed one at a time, for both classifiers are as given below.

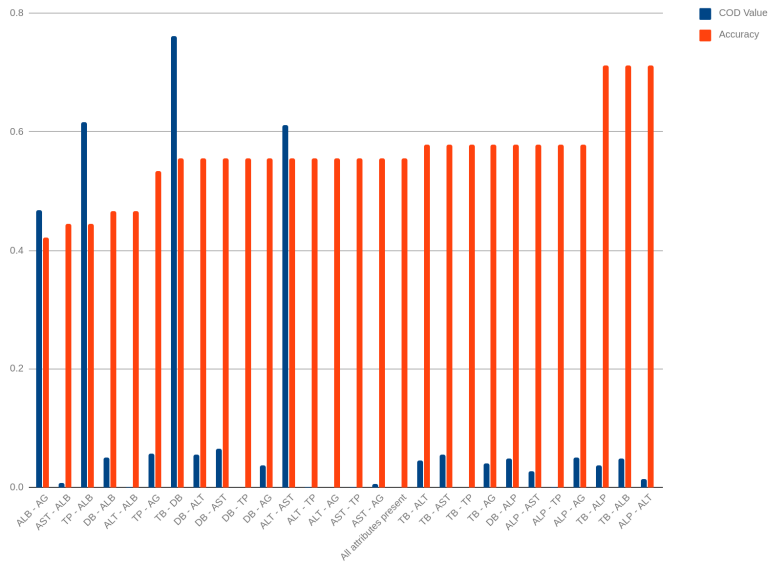

Figure 4: Relationship between COD and Accuracy for all Attribute pairs (Iterative Dichotomiser 3)

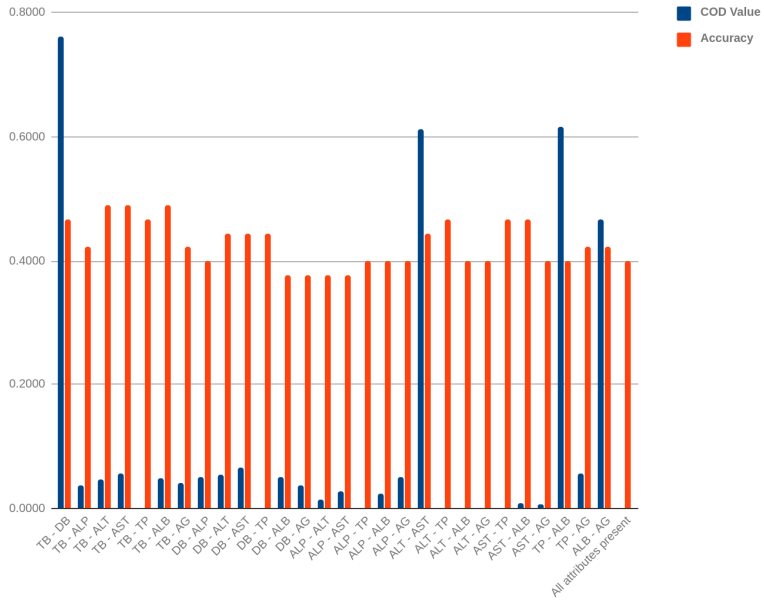

Figure 5: Relationship between COD and Accuracy for all Attribute pairs (Naïve Bayes Classifier)

4.3. Attribute Evaluator Ranker

4.3.1 Gini Index

The Gini index was calculated for all the attributes. From the figure 6, given below, it can be seen that the attributes TP and ALB had the highest values, while ALP and ALT had the lowest values.

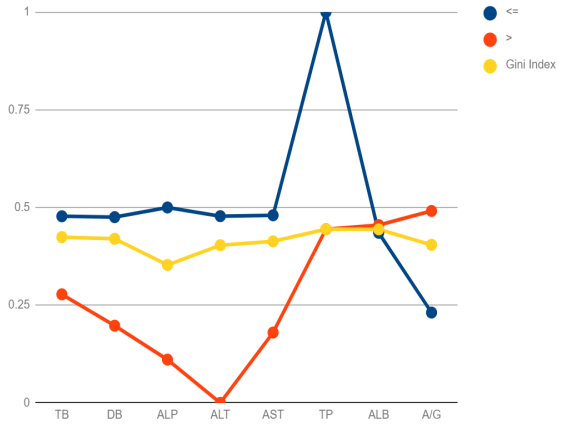

Figure 6: Gini Index Graphical Analysis

4.3.2 Information Gain

From the figure 7, out of all the attributes; ALP, ALT, AST and AG shows the highest information gain. These attributes can be used to obtain a better accuracy.

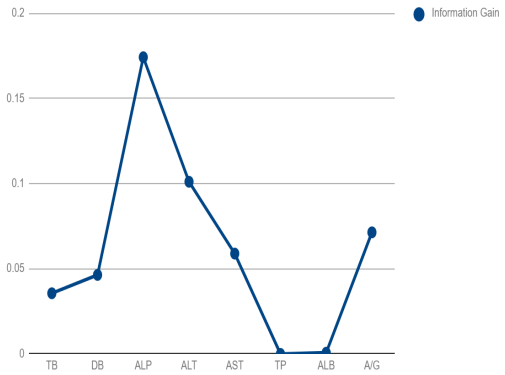

Figure 7: Information Gain Graphical Analysis

4.3.3 Gain Ratio

The gain ratio results are the same as that for the information gain in Figure 8, The attributes ALP, ALT, AST and AG corresponds to the highest gain ratio values.

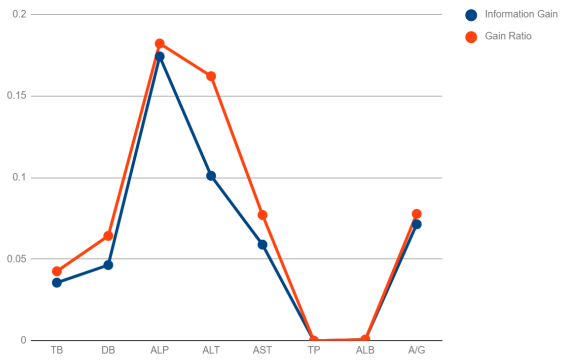

Figure 8: Gain Ratio Graphical Analysis

4.3.4 One- R

The one-r method gives the error rate of all the attributes. The attributes DB, ALT, TP have low error rates as seen in figure 9, compared to the remaining attributes. Hence these attributes can be used to give better results with less error rates.

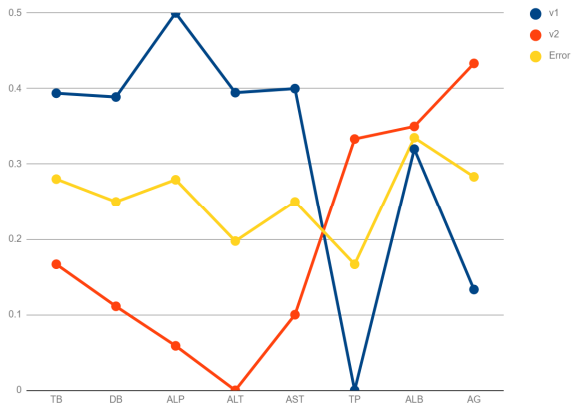

Figure 9: One R/Error Graphical Analysis

5. CONCLUSIONS

5.1. Conclusions

The system which is used for diagnosing the proper functioning of the liver may consist of some errors as mentioned in the chapters above. The proposed system intended the error rate of falsely diagnosing to be reduced to a minimum. An initial model to test the best discretisation method gave the result that brute force gave the best output. Coefficient of Determination were also computed. It was determined that the classifier “Iterative Dichotomiser 3” gives better, faster and more accurate results as compared to the classifier “Naïve Bayes”. Through the Attribute Evaluator Rankers like Information Gain, Gini Index, Gain Ration and One-R, each of the attributes was ranked and the best attributes were determined.

5.2. Future Scope

The two major steps in the dependency guided wrapper is the attribute selection search and the attribute evaluator rankers. The attribute evaluator rankers have been discussed earlier. The improvements to the approach could be the use of Pearson’s correlation coefficient over Coefficient of Determination. The alternative approach to that of the wrapper, for finding the best attribute subset is the filter approach. The features, or attributes, are selected based upon a step in pre-processing. However, unlike the wrapper approach, the filter approach does not take into consideration, the classifier, but works independent of the same. Another substitute for wrapper can be the embedded approach, where a combination of a filter as well as a wrapper can be used. Regularisation techniques may be used for the same.

REFERENCES

1. *Cortizo, J. C., et al (2007, December). Wrapping the naive bayes classifier to relax the effect of dependencies. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 229-239). Springer, Berlin, Heidelberg.*
2. *Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.*
3. *Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 21(1).*
4. *Abdullah, S. B., Man, Z., Ismail, L., Maulud, A., & Bustam, M. A. (2013). Ionic liquids classification for fuel desulphurization. Int. J. Gen. Eng. Technol, 2, 29-38.*
5. *J Han, et al.(2011). Data mining: concepts and techniques. Elsevier.*
6. *J K Sharma.(2012). Business Statistics: Problems & Solutions. Vikas Publishing House.*